arXiv:2601.06453v1 [cs.AI] 10 Jan 2026

ConSensus: Multi-Agent Collaboration for Multimodal Sensing

Hyungjun Yoon[1*] **Mohammad Malekzadeh**[2]
Sung-Ju Lee[1] **Fahim Kawsar**[2,3] **Lorena Qendro**[2]

[1]KAIST [2]Nokia Bell Labs [3]University of Glasgow

Abstract

Large language models (LLMs) are increasingly grounded in sensor data to perceive and reason about human physiology and the physical world. However, accurately interpreting heterogeneous multimodal sensor data remains a fundamental challenge. We show that a single monolithic LLM often fails to reason coherently across modalities, leading to incomplete interpretations and prior-knowledge bias. We introduce **ConSensus**, a training-free multi-agent collaboration framework that decomposes multimodal sensing tasks into specialized, *modality-aware agents*. To aggregate agent-level interpretations, we propose a *hybrid fusion* mechanism that balances *semantic* aggregation, which enables cross-modal reasoning and contextual understanding, with *statistical* consensus, which provides robustness through agreement across modalities. While each approach has complementary failure modes, their combination enables reliable inference under sensor noise and missing data. We evaluate ConSensus on five diverse multimodal sensing benchmarks, demonstrating an average accuracy improvement of 7.1% over the single-agent baseline. Furthermore, ConSensus matches or exceeds the performance of iterative multi-agent debate methods while achieving a 12.7× reduction in average fusion token cost through a single-round hybrid fusion protocol, yielding a robust and efficient solution for real-world multimodal sensing tasks.

1 Introduction

Large language models (LLMs) are being extended beyond text to perceive and reason about the physical world and human physiology, motivated by emerging applications in embodied interactions (Driess et al., 2023) and health monitoring (Kim et al., 2024). This requires LLMs to be systematically grounded in sensor data that encodes underlying motions, biosignals, and environmental measurements. Recent studies have explored diverse grounding strategies, including transforming raw signals into descriptive or visual prompts (Liu et al., 2023; Yoon et al., 2024) and jointly aligning learned sensor encoders with LLM representations (Zhang et al., 2025). In contrast to traditional deep learning for sensing (Radu et al., 2018), integration with LLMs enables the resolution of diverse sensing tasks in an interpretable, training-free manner by utilizing the models' internal world knowledge.

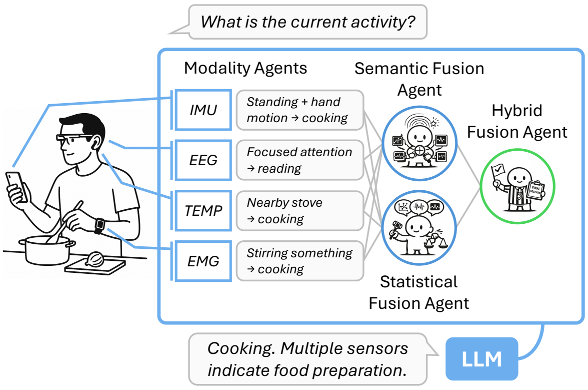

What is the current activity?
Modality Agents
Semantic Fusion Agent
Hybrid Fusion Agent
IMU
Standing + hand motion → cooking
EEG
Focused attention → reading
TEMP
Nearby stove → cooking
EMG
Stirring something → cooking
Statistical Fusion Agent
Cooking. Multiple sensors indicate food preparation.
LLM

Figure 1: Illustration of ConSensus. Modality-aware agents provide specialized interpretations aggregated via hybrid fusion for multimodal sensing.

Physical tasks are inherently multimodal, with distinct and complementary aspects captured by heterogeneous sensors (Radu et al., 2018). As a result, multimodal sensing is crucial for combining complementary information across modalities. It becomes particularly important in real-world settings where individual sensors might be unreliable or missing. However, this necessity introduces the challenge of *sensor fusion*: how to systematically integrate heterogeneous modalities into a coherent decision process. Despite growing interest in multimodal LLMs, it remains unclear how LLMs should reason over sensor-specific representations and aggregate their semantic interpretations.

---

[*] Work done during the author's internship at Nokia Bell Labs. Email: hyungjun.yoon@kaist.ac.kr.

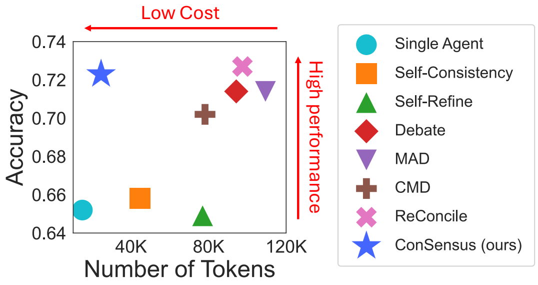


Figure 2: ConSensus (top-left) achieves higher accuracy (y-axis) at lower cost (x-axis) compared to baselines.

To address the challenge, we leverage multi-agent collaboration as a mechanism for decomposing multimodal reasoning. Recent studies have assigned specialized roles to multiple LLM instances to generate diverse reasoning paths for solving complex tasks (Li et al., 2023). However, existing multi-agent frameworks primarily focus on diversifying textual reasoning, and do not address the unique challenges introduced by heterogeneous sensors with asymmetric reliability. In this work, we propose a multi-agent framework for multimodal sensing that explicitly encourages modality-specific agents to produce distinct reasoning trajectories and integrates their semantic interpretations via a structured collaboration protocol.

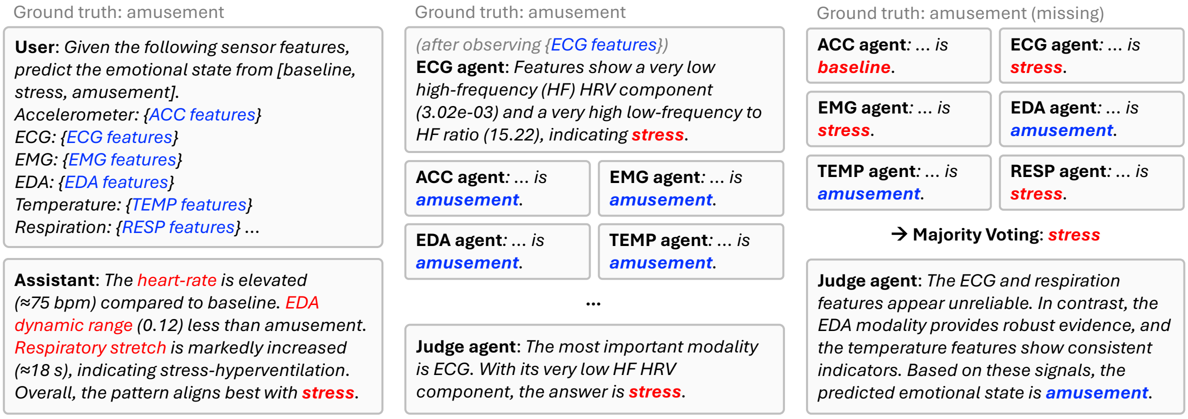
Our design is guided by three empirically grounded observations. First, we observe that a single LLM struggles to jointly ground its reasoning across all modalities, often producing incomplete interpretations when faced with heterogeneous sensor inputs. This limitation motivates partitioning agents by modality, thereby encouraging independent and complementary reasoning paths. Second, during modality fusion, LLM-based semantic aggregation exhibits pronounced prior-knowledge bias, in which errors from particular modalities are disproportionately amplified and propagated to the final decision. Third, although statistical aggregation mechanisms such as majority voting can mitigate semantic bias, they break down in realistic sensing settings, where missing or unreliable sensor inputs can distort voting outcomes. Taken together, these observations expose a fundamental trade-off between semantic and statistical fusion, motivating a collaboration protocol that balances their complementary strengths and failure modes.

Based on these insights, we present **ConSensus**, a training-free multi-agent framework for multimodal sensing. We define *modality agents* as LLM instances that operate under modality-specific roles and generate independent semantic interpretations of the same task. To aggregate these interpretations into a final decision, we propose a principled *hybrid fusion* strategy. A *semantic fusion agent* performs reasoning-level aggregation to integrate the semantic outputs of modality agents and produce an initial prediction, while a *statistical fusion agent* anchors its decision to a majority-voted outcome across modality agents, providing complementary perspective against prior-driven semantic bias. A final *hybrid fusion agent* is instructed to jointly observe the semantic and statistical fusion outputs to produce the final prediction as a coordinator role.

We evaluate ConSensus across five diverse multimodal sensing tasks, covering a broad range of heterogeneous sensor modalities. Experimental results demonstrate that ConSensus consistently achieves a 7.1% accuracy improvement over the single-agent baseline. Ablation studies further show that the proposed hybrid fusion mechanism effectively navigates the trade-off between semantic and statistical fusion, adaptively favoring the more reliable reasoning path under varying sensor conditions. Compared to state-of-the-art multi-agent debate baselines applied to the same modality agents, ConSensus matches or outperforms accuracy while reducing average fusion token cost by 12.7× (Figure 2). This efficiency directly follows from our architectural design: while existing multi-agent frameworks depend on costly iterative debates to refine predictions, ConSensus employs a single-round, structured fusion protocol, achieving substantially lower inference overhead without sacrificing accuracy.

Our primary contributions are as follows:

- We propose ConSensus, the first multi-agent framework for multimodal sensing that decomposes heterogeneous sensor inputs into specialized, modality-aware agents. ConSensus operates without additional training, enabling direct deployment across diverse tasks.
- We present a hybrid fusion strategy that balances the trade-off between semantic aggregation and statistical consensus under sensor uncertainty.
- We evaluate ConSensus across five diverse multimodal sensing tasks, demonstrating an accuracy gain of 7.1% over the single-agent baseline.
- We show that our single-round fusion protocol matches or outperforms state-of-the-art multi-agent debate methods while achieving a 12.7× reduction in fusion token cost, substantially improving inference efficiency and scalability.


Ground truth: amusement

User: *Given the following sensor features, predict the emotional state from [baseline, stress, amusement]. Accelerometer: {ACC features} ECG: {ECG features} EMG: {EMG features} EDA: {EDA features} Temperature: {TEMP features} Respiration: {RESP features} ...*

Assistant: *The heart-rate is elevated (≈75 bpm) compared to baseline. EDA dynamic range (0.12) less than amusement. Respiratory stretch is markedly increased (≈18 s), indicating stress-hyperventilation. Overall, the pattern aligns best with* ***stress****.*

Ground truth: amusement

(after observing {ECG features})
ECG agent: *Features show a very low high-frequency (HF) HRV component (3.02e-03) and a very high low-frequency to HF ratio (15.22), indicating* ***stress****.*

ACC agent: *... is* ***amusement****.*

EMG agent: *... is* ***amusement****.*

EDA agent: *... is* ***amusement****.*

TEMP agent: *... is* ***amusement****.*

...

Judge agent: *The most important modality is ECG. With its very low HF HRV component, the answer is* ***stress****.*

Ground truth: amusement (missing)

ACC agent: *... is* ***baseline****.*

ECG agent: *... is* ***stress****.*

EMG agent: *... is* ***stress****.*

EDA agent: *... is* ***amusement****.*

TEMP agent: *... is* ***amusement****.*

RESP agent: *... is* ***stress****.*

→ **Majority Voting**: ***stress***

Judge agent: *The ECG and respiration features appear unreliable. In contrast, the EDA modality provides robust evidence, and the temperature features show consistent indicators. Based on these signals, the predicted emotional state is* ***amusement****.*

(a) Single-agent inference. (b) Modality agents with a judge agent. (c) Data with 30% missing modalities.

Figure 3: Examples of LLM-based multimodal sensing on WESAD (Schmidt et al., 2018) using `gpt-oss-20B`.

2 Background and Motivation

2.1 Problem Formulation

We address the *multimodal sensing task*, in which LLMs are leveraged as training-free reasoning engines across diverse tasks to produce semantically interpretable outputs. A multimodal sensing task requires integrating heterogeneous sensor modalities to produce a single, coherent inference.

Formally, given N modalities with inputs $M = \{m_1, m_2, \ldots, m_N\}$ and a task description T, the objective is to predict an output $y \in \mathcal{Y}$ that solves the task by jointly reasoning over all available modalities. In this work, each modality input m_i is represented using standard, hand-crafted sensor features (e.g., summary statistics such as `mean` and `std`), while noting that our formulation is agnostic to the specific input representation and can be readily extended to alternative forms such as raw signals, images, or learned embeddings (Liu et al., 2023; Yoon et al., 2024; Zhang et al., 2025).

2.2 Motivating Examples

A straightforward instantiation is to present all modality features (M) jointly to a single LLM agent together with the task description T. Figure 3a illustrates an affective state inference task posed to a single agent using multiple sensor modalities (Schmidt et al., 2018). While the agent generates plausible interpretations for some modalities, evidence from other sensors is frequently overlooked, resulting in an incomplete reasoning process and, ultimately, an incorrect prediction. We attribute this failure to *context overload* and *modality dominance* within a single agent, where subtle but critical modality-specific evidence is overridden by salient prominent signals.

Observation 1: A single agent often produces an *incomplete* cross-modal interpretation in multimodal sensing tasks.

To ensure completeness, we decompose multimodal interpretation into *modality agents*, each of which follows an independent reasoning path conditioned on a specific sensor modality (m_i). This decomposition produces multiple semantic interpretations for the same task, which must subsequently be aggregated into a single decision. Common aggregation strategies include (i) employing an additional *judge agent* (Liang et al., 2024) to synthesize a final decision from all agent outputs, or (ii) applying majority voting (Du et al., 2024) over individual predictions.

However, we find that the judge agent often relies on prior domain knowledge to disproportionately weight certain modalities (e.g., privileging clinically salient signals such as ECG), leading to incorrect decisions when those modalities are unreliable (Figure 3b). In real-world deployments, sensor signals often deviate from canonical patterns due to device-specific characteristics or environmental noise. While complementary modalities with consistent evidence can mitigate such noise, the LLM's *knowledge bias* frequently causes error propagation; incorrect assumptions associated with a perceived "important" modality override correct evidence from other modalities.

Observation 2: An LLM-based judge is prone to *knowledge bias*, whereby prior domain knowledge can mislead consensus even when robust evidence exists in other modalities.

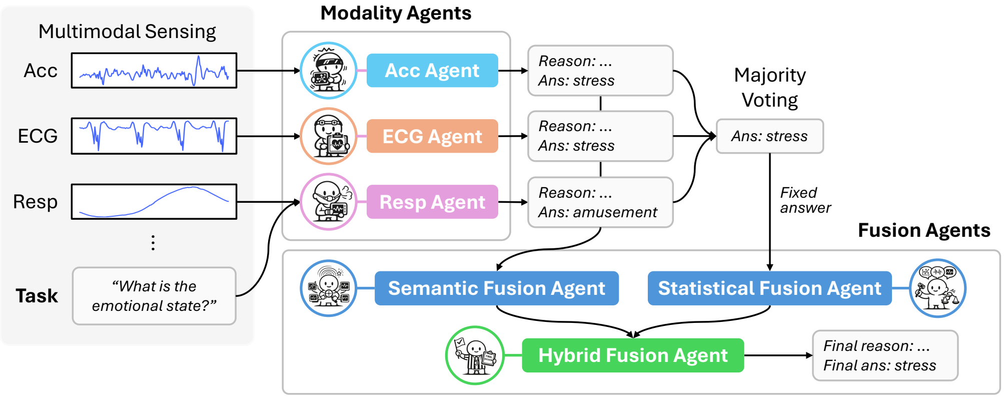


Figure 4: Overview of ConSensus: (i) Modality agents generate specialized per-sensor interpretations; (ii) a semantic fusion agent aggregates cross-modal reasoning; (iii) a statistical fusion agent provides an output that anchors the reasoning to the majority; and (iv) a hybrid fusion agent reconciles both outputs to yield the final decision.

A natural approach to mitigating knowledge bias is majority voting, which relies on statistical aggregation rather than semantic reasoning or prior knowledge. According to Condorcet's Jury Theorem (CJT) (Condorcet, 1785), if each agent maintains a reliability > 0.5, the collective accuracy provably converges to 100% as the number of agents increases. This suggests that, in theory, a sufficient number of reliable modalities can yield high accuracy through simple majority voting.

Conversely, when modalities are missing, corrupted, or systematically degraded (i.e., *sensor failure*), a common scenario in real-world sensing (Du et al., 2020), majority voting can fail catastrophically. Figure 3c illustrates such a failure case under 30% missing modalities. In these settings, sensor failures violate the core assumption of CJT (i.e., voter reliability < 0.5), causing statistical aggregation to converge toward an incorrect consensus. In contrast, the judge agent identifies uncertainty cues to discount unreliable reasoning and prioritize complementary modalities that provide consistent rationales.

> **Observation 3:** Majority voting is prone to *sensor failure*, often producing unreliable votes and false consensus in multimodal sensing.

Together, observations 1–3 reveal complementary failure modes in multimodal reasoning: single-agent reasoning suffers from incomplete cross-modal interpretation (Observation 1), semantic aggregation is vulnerable to knowledge bias (Observation 2), and statistical aggregation is fragile under sensor failure (Observation 3). These observations jointly expose a fundamental tension between semantic-level and statistical aggregation.

3 Method

We introduce **ConSensus**, a training-free Conversational Sensor Fusion framework that performs multimodal sensing via structured, role-specialized multi-agent collaboration (Figure 4). ConSensus comprises four defined agent roles: (i) ***modality agents***, each specialized to a single sensor modality and responsible for producing task-relevant semantic interpretations; (ii) a ***semantic fusion agent*** that integrates modality-agent outputs into a holistic prediction by reasoning over cross-modal semantic evidence; (iii) a ***statistical fusion agent*** that reasons over a fixed prediction anchored to a majority-vote consensus of modality-level predictions, providing robustness against prior-driven knowledge bias; and (iv) a ***hybrid fusion agent*** that arbitrates between the outputs of the semantic and statistical fusion agents. By explicitly balancing knowledge-grounded and consensus-grounded reasonings, the hybrid fusion agent produces the final prediction.

3.1 Modality Agents

ConSensus instantiates a set of **modality agents** $\{A_1^{\text{mod}}, A_2^{\text{mod}}, \ldots, A_N^{\text{mod}}\}$, where each agent A_i^{mod} is assigned to a single sensor modality m_i and is responsible for producing a task-relevant semantic interpretation of that modality under the query T. Each agent receives a modality-conditioned prompt (m_i, T) and outputs (i) a modality-specific prediction $\hat{y}_i \in \mathcal{Y}$ and (ii) a rationale r_i that explicitly grounds the prediction in evidence. This modularization isolates modality-level reasoning, reducing cross-modal interference within the prompt and preserving complementary evidence that may otherwise be diluted under joint prompting.

3.2 Fusion Agents

To reach a final prediction, ConSensus aggregates the outputs of the modality agents, $\{(\hat{y}_i, r_i)\}_{i=1}^{N}$. As discussed in Section 2, semantic and statistical aggregation paradigms exhibit complementary strengths and distinct failure modes. To balance these trade-offs, we define a *hybrid fusion* approach comprising three fusion agents.

Semantic Fusion Agent ($A^{\text{fuse}}_{\text{sem}}$). It implements a *knowledge-grounded aggregation* that synthesizes semantic evidence across modality-agent outputs $\{(\hat{y}_1, r_1), \ldots, (\hat{y}_N, r_N)\}$ using the LLM's prior knowledge. It represents a semantic inductive bias that emphasizes cross-modal coherence, causal plausibility, and high-level domain knowledge when forming an aggregated hypothesis.

Statistical Fusion Agent ($A^{\text{fuse}}_{\text{stat}}$). It implements a *consensus-grounded aggregation* by anchoring its reasoning to a fixed majority-voted prediction $\hat{y}_{\text{vote}} = \arg\max_{c \in \mathcal{Y}} \sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = c]$. Rather than generating a new prediction, it produces a rationale that justifies this statistical consensus, representing a complementary inductive bias that prioritizes robustness to individual agent errors and suppresses prior-driven knowledge bias.

Hybrid Fusion Agent ($A^{\text{fuse}}_{\text{hyb}}$). We formalize hybrid fusion as an instance-wise arbitration between two complementary inductive biases: $(\hat{y}_{\text{sem}}, r_{\text{sem}}) = A^{\text{fuse}}_{\text{sem}}(\{(\hat{y}_i, r_i)\}_{i=1}^{N})$ and $(\hat{y}_{\text{stat}}, r_{\text{stat}}) = A^{\text{fuse}}_{\text{stat}}(\{(\hat{y}_i, r_i)\}_{i=1}^{N})$. The hybrid fusion agent $A^{\text{fuse}}_{\text{hyb}}$ performs instance-wise arbitration: $(\hat{y}, r) = A^{\text{fuse}}_{\text{hyb}}\big((\hat{y}_{\text{sem}}, r_{\text{sem}}), (\hat{y}_{\text{stat}}, r_{\text{stat}})\big)$ which selects between two complementary aggregation hypotheses grounded in (i) semantic reasoning and (ii) statistical consensus as a *coordinator* role.

Unlike prior methods that rely only on knowledge-grounded reasoning, the hybrid fusion agent is additionally exposed to statistically grounded reasoning. This design introduces a complementary, data-driven perspective into the decision process, allowing ConSensus to maintain robustness under sensor failure via semantic coherence, while simultaneously mitigating knowledge bias by taking account of statistical stability. Full prompt templates are provided in Appendix C.

4 Experiments

4.1 Setup

Datasets. We evaluate ConSensus across five diverse multimodal sensing benchmarks, spanning both physiological and physical activity recognition tasks: (i) WESAD (Schmidt et al., 2018) for affective state recognition; (ii) SleepEDF (Kemp et al., 2000) for sleep stage classification; (iii) ActionSense (DelPreto et al., 2022) for recognizing four categories of kitchen activities (e.g., peeling, washing); (iv) MMFit (Strömbäck et al., 2020) for gym exercise recognition; and (v) PAMAP2 (Reiss and Stricker, 2012) for daily activity recognition. The datasets cover a wide range of 12 distinct sensor modalities collected from different locations and devices. Raw sensor data were processed into standard, hand-crafted feature representations following established protocols (Schmidt et al., 2018; Xu et al., 2025). We detail the sensor modalities, preprocessing, and splits in Appendix A.

Prompts. We employ text-only prompts that embed extracted sensor features within a structured prompt template. Each prompt includes a detailed description of the target task and data, derived directly from the original dataset documentation. Furthermore, we employ a 1-shot in-context learning strategy, providing one representative example per class to enhance the model reasoning. We provide the prompt details in Appendix C.

Models. We use `gpt-oss-20B` as the main backbone (OpenAI et al., 2025). Moreover, we evaluate scale and architecture variations using `gpt-oss-120B`, `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) and `Llama-4-Scout-17B-16E` (Meta AI, 2025). We use a temperature of 0 to ensure deterministic and reproducible outputs.

Baselines. We evaluate ConSensus against three representative baselines: (i) **Single-Agent**, where all modality features are concatenated into a single prompt, (ii) **Self-Consistency** (SC) (Wang et al., 2023), which samples multiple reasoning paths and selects the most consistent prediction (with temp= 0.7), and (iii) **Self-Refine** (SR) (Madaan et al., 2023), which iteratively generates feedback and refine an initial response. For SC, we use three LLM instances, and for SR, we perform two refinement steps to ensure a comparable inference budget with ConSensus, which involves two intermediate fusion steps prior to the final hybrid decision.

To further evaluate the effectiveness of our hybrid fusion design, we compare ConSensus against state-of-the-art multi-agent debate frameworks adapted to operate over the same set of modality agents: (i) **Debate** (Du et al., 2024), where agents iteratively refine responses by observing the outputs of other agents, with the final decision

Method	WESAD	SleepEDF	ActionSense	MMFit	PAMAP2	Avg.
Single-agent baselines						
Single-Agent	0.793 ± 0.033	0.519 ± 0.031	0.577 ± 0.032	0.819 ± 0.022	0.551 ± 0.027	0.652 ± 0.027
+ Self-Consistency	0.786 ± 0.035	0.541 ± 0.031	0.555 ± 0.031	0.862 ± 0.019	0.547 ± 0.027	0.658 ± 0.027
+ Self-Refine	0.747 ± 0.035	0.551 ± 0.031	0.566 ± 0.031	0.822 ± 0.022	0.563 ± 0.026	0.650 ± 0.026
Modality agents + multi-agent debate baselines, requiring **76K additional tokens** per sample						
+ Debate	0.873 ± 0.027	0.548 ± 0.032	0.609 ± 0.031	**0.984** ± 0.007	0.561 ± 0.028	0.715 ± 0.028
+ MAD	0.847 ± 0.029	0.562 ± 0.031	0.613 ± 0.031	0.960 ± 0.011	**0.589** ± 0.026	0.714 ± 0.026
+ CMD	0.840 ± 0.030	0.578 ± 0.031	0.589 ± 0.034	0.962 ± 0.011	0.539 ± 0.025	0.702 ± 0.025
+ ReConcile	0.880 ± 0.027	0.571 ± 0.030	**0.640** ± 0.031	0.964 ± 0.011	0.579 ± 0.026	**0.727** ± 0.026
Modality agents + fusion agents (ours), requiring **6K additional tokens** per sample						
+ Semantic Fusion	0.825 ± 0.031	0.580 ± 0.031	0.605 ± 0.033	0.964 ± 0.011	0.559 ± 0.026	0.707 ± 0.026
+ Statistical Fusion	**0.927** ± 0.021	0.592 ± 0.032	0.597 ± 0.033	0.960 ± 0.011	0.534 ± 0.026	0.722 ± 0.026
ConSensus	0.880 ± 0.029	**0.600** ± 0.031	0.611 ± 0.031	0.967 ± 0.010	0.558 ± 0.026	0.723 ± 0.026

Table 1: Comparison of ConSensus with baseline methods. Ablation results for the individual semantic and statistical fusion agents are reported for comparison. The best and second-best accuracies are **bold** and underline.

determined by majority voting; (ii) **MAD** (Liang et al., 2024), which follows a similar protocol to Debate but employs an judge agent to steer the final decision; (iii) **CMD** (Wang et al., 2024), where agents are partitioned into groups (we use two groups) such that agents within each group share full responses, while only prediction counts are exchanged across groups; and (iv) **ReConcile** (Chen et al., 2024), where agents output an explicit confidence score and arrive at a final decision via confidence-weighted voting. As SC and SR, we fix the number of debate rounds to two. We emphasize that, while these baselines are designed for multi-round debates, ConSensus performs aggregation in a single round, highlighting a fundamental difference in both efficiency and interaction structure.

Importantly, ConSensus is a *training-free* framework. By leveraging the world knowledge and reasoning capabilities of pre-trained LLMs, our method bypasses the need for large-scale data collection and task-specific model training. Consequently, we do not include traditional learning-based sensor fusion methods (Radu et al., 2018) as baselines, since they rely on supervised training over substantial task-specific data and are therefore not directly comparable to a training-free setting.

4.2 Results

Effect of Modality Agents. We first evaluate the impact of modality agents by comparing ConSensus against single-agent reasoning baselines. As shown in Table 1, ConSensus consistently outperforms the Single-Agent baseline by an average accuracy margin of 7.1%. Notably, even without hybrid fusion, both semantic-only and statistical-only fusion variants achieve substantial performance gains, indicating that modality-specific agent decomposition alone yields strong improvements. While Self-Consistency yields only marginal improvements (an average gain of 0.6%), its performance remains substantially below that of ConSensus. Qualitative analysis (Appendix D) shows that the Single-Agent baseline often omits modality-specific interpretations, consistent with Observation 1. In contrast, ConSensus ensures complete modality coverage, resulting in more reliable and well-grounded decisions.

Effect of Hybrid Fusion. Table 1 further compares ConSensus with semantic-only and statistical-only fusion variants. While statistical fusion outperforms semantic fusion by an average margin of 1.5%, the optimal fusion strategy varies substantially across datasets. Semantic fusion performs best when sensor data align well with LLM prior knowledge (e.g., PAMAP2, 2.5%), whereas statistical fusion dominates under unexpected or noisy data characteristics (e.g., WESAD, 10.2%). This is consistent with Observation 2: in WESAD, the semantic fusion agent frequently overweights incorrect ECG-derived predictions due to knowledge bias, whereas majority voting aggregates correct evidence from secondary modalities.

Importantly, the hybrid fusion agent resolves this trade-off, achieving the highest average accuracy across datasets. On SleepEDF, ActionSense, and MMFit, the hybrid agent surpasses both semantic-only and statistical-only fusion agents. This demonstrates the hybrid agent's ability to follow semantic reasoning when it is reliable and to revert to statistical consensus when knowledge bias would otherwise lead to incorrect decisions.

Comparison with Multi-Agent Debate. We fur-

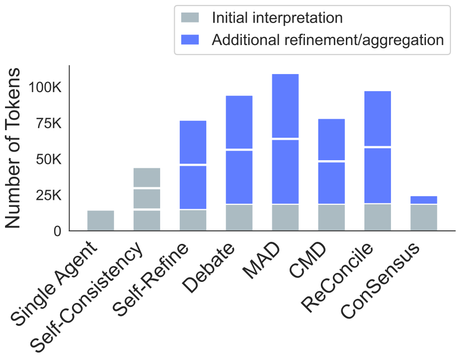


Figure 5: Average input tokens per inference across datasets. Gray bars denote tokens required for initial interpretation, and blue bars denote aggregation or refinement tokens, segmented by rounds.

ther compare ConSensus with iterative multi-agent debate baselines built on the same set of modality agents. ConSensus achieves the highest accuracy among all baselines, with the exception of ReConcile, which attains comparable performance (72.7% vs. 72.3%). Notably, all debate baselines rely on multi-round interactions, incurring substantial token overhead. In contrast, ConSensus achieves comparable or superior accuracy using a single-round fusion protocol, resulting in up to 12.7× reduction in fusion tokens compared to the debate baselines. We further evaluate non-iterative variants of the debate baselines in Appendix B, showing that ConSensus consistently outperforms debate baselines under similar number of tokens.

Token Efficiency. Figure 5 reports the average token usage per inference across all datasets. Gray bars represent tokens consumed during initial interpretation, including system prompts and sensor feature descriptions required to derive an initial prediction. These costs remain consistent across all methods, as modality agents partition the same underlying sensory information rather than introducing additional content. Blue bars represent tokens consumed by refinement, debate, or fusion procedures. While ReConcile, one of the strongest baselines, requires 78.6K tokens for aggregation per inference to achieve accuracy comparable to ConSensus, ConSensus requires only 6K tokens. Averaged across all multi-agent debate baselines, ConSensus reduces aggregation tokens by 12.7×. This result highlights the substantial token efficiency of ConSensus, achieved through its single-round structured fusion design.

Robustness to Sensor Failure. We simulate sensor failure by randomly omitting 10%, 30%, and 50% of sensor modalities. As shown in Figure 6, ConSensus maintains a consistent performance advantage, outperforming the Single-Agent baseline by an average of 7.1%, 4.5%, 8.2%, 1.2% across the respective missingness ratios.

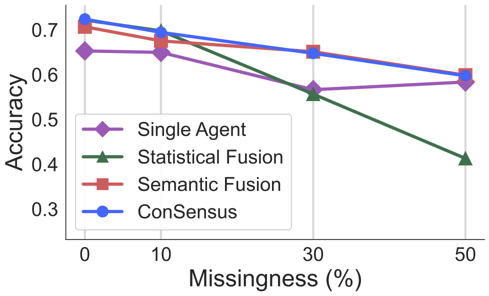


Figure 6: Accuracy under random modality omission at varying missingness levels. ConSensus maintains higher accuracy by adaptively balancing semantic and statistical fusion.

We observe that the statistical fusion degrades sharply as sensor missingness increases, collapsing to 41.4% at 50% missingness. This confirms Observation 3: majority voting is fragile when anchored to a high ratio of unreliable votes. In contrast, semantic fusion exhibits remarkable resilience, maintaining 59.9% accuracy even in extreme failure scenarios. Importantly, ConSensus's hybrid fusion dynamically prioritizes the semantic reasoning as statistical certainty drops. Consequently, ConSensus outperforms the statistical fusion by 9.1% at 30% and 18.4% at 50% missingness, effectively mitigating catastrophic degradation by selectively following the semantic fusion agent.

Effect of LLM. We evaluate ConSensus across diverse open-sourced LLMs, including `gpt-oss-120B`, `Llama-3.1-8B-Instruct`, and the Mixture-of-Experts (MoE) based `Llama-4-Scout-17B-16E` (109B total parameters). Table 2 compares ConSensus against the Single-Agent baseline, the best-performing debate baseline ReConcile, and the semantic and statistical fusion agents.

Across all LLMs, ConSensus consistently outperforms the Single-Agent baseline and exhibits robust hybrid fusion behavior, effectively balancing semantic and statistical fusion results. For `gpt-oss-120B`, semantic and statistical fusion achieve comparable performance, and ReConcile outperforms ConSensus, suggesting that the prior knowledge in larger parameters benefit the fusion and further debates. In contrast, on `Llama` models, ConSensus consistently outperforms ReConcile while requiring fewer tokens (up to 13.1×), indicating that ReConcile's advantages are not robust across LLMs. Notably, on the small `Llama-3.1-`

Method	WE-SAD	Sleep EDF	Action Sense	MM-Fit	PA-MAP2	Avg.
`gpt-oss-120B`						
Single agent						
Single-Agent	0.807	0.580	0.564	0.730	0.592	0.654
Modality agents						
+ ReConcile	**0.887**	**0.588**	**0.604**	**0.980**	**0.636**	**0.739**
+ Semantic Fusion	0.793	0.572	0.600	0.967	0.633	0.713
+ Statistical Fusion	0.880	0.580	0.572	0.960	0.572	0.713
ConSensus	0.833	0.572	0.584	0.967	0.611	0.713
`Llama-3.1-8B-Instruct`						
Single agent						
Single-Agent	0.493	0.292	0.288	0.167	0.225	0.293
Modality agents						
+ ReConcile	0.673	0.304	0.296	0.390	0.231	0.379
+ Semantic Fusion	0.773	0.368	0.288	0.493	0.344	0.453
+ Statistical Fusion	**0.787**	**0.416**	**0.324**	**0.543**	0.342	**0.482**
ConSensus	0.780	0.368	0.296	0.487	**0.347**	0.456
`Llama-4-Scout-17B-16E`						
Single agent						
Single-Agent	0.660	0.416	0.268	0.293	0.386	0.405
Modality agents						
+ ReConcile	0.853	0.556	0.572	0.877	0.497	0.671
+ Semantic Fusion	0.833	0.536	0.588	0.900	0.525	0.676
+ Statistical Fusion	**0.860**	**0.568**	**0.604**	**0.910**	0.531	**0.695**
ConSensus	**0.860**	0.548	0.592	0.903	**0.542**	0.689

Table 2: Performance comparison of Single-Agent baseline, ReConcile, Semantic Fusion, Statistical Fusion, and ConSensus across four LLM backbones.

`8B-Instruct` model, the Single-Agent baseline performs poorly (29.3%), and ReConcile yields limited improvement (+8.6%), whereas ConSensus achieves a substantially larger gain (+16.3%). This highlights the strong potential of ConSensus to enhance multimodal sensing capabilities for smaller, resource-constrained LLMs. Overall, ConSensus delivers consistent performance gains across model families while maintaining high token efficiency.

5 Related Work

5.1 LLM-based Sensing

Grounding LLMs with sensor data enables a broad spectrum of sensing tasks by leveraging world knowledge and semantic reasoning to interpret rich real-world contexts. Early works (Xu et al., 2024; Liu et al., 2023) demonstrated that LLMs can interpret raw or processed sensor data to solve real-world sensing problems. Subsequent studies have explored broader reasoning capabilities over sensor inputs (Ouyang and Srivastava, 2024) , conversational analysis over long-term sensor traces (Yu et al., 2025). To improve sensor understanding, recent works proposed transforming sensor data into visual representations (Yoon et al., 2024), aligning LLMs with modality-specific encoders (Imran et al., 2025), or pre-training large-scale sensor language models (Zhang et al., 2025; Langer et al., 2025; Chan et al., 2024). Additionally, agentic approaches have been proposed to decompose complex sensor analysis into modular actions or workflows (Shen et al., 2025; Heydari et al., 2025).

However, most existing studies focus on single or a small number of sensor modalities and do not explicitly address the challenges of heterogeneous sensors. In this work, we investigate multimodal sensing with LLMs, identifying fundamental challenges related to dense inputs and the absence of effective semantic-level fusion mechanisms, and propose a solution to address these limitations.

5.2 Multi-Agent Collaboration

Multi-agent collaboration enables complex task execution by assigning specialized roles and coordinating interactions among multiple agents (Li et al., 2023; Hong et al., 2024). Prior work has demonstrated the effectiveness of multi-agent collaboration in coding (Qian et al., 2024), scientific research (Lu et al., 2024), and tool use (Wu et al., 2024). Recent works employ multi-agent debate mechanisms to improve reasoning accuracy by leveraging iterative deliberation to converge on more robust solutions (Du et al., 2024; Liang et al., 2024; Chen et al., 2024; Wang et al., 2024). Nevertheless, most existing frameworks are designed for text-centric benchmarks and task domains such as writing and coding, rather than sensor-driven or multimodal settings. In this work, we extend the scope of multi-agent collaboration to multimodal sensing by introducing a role-specialized framework tailored to heterogeneous sensor inputs.

6 Conclusion

We propose a multi-agent collaboration framework for multimodal sensing that decomposes sensing tasks into modality-aware agents, enabling independent interpretation of heterogeneous sensor streams. Our hybrid fusion mechanism navigates the tradeoff between semantic aggregation and statistical consensus of the modality-specific outputs. While semantic reasoning remains robust under sensor failures, the statistical anchor mitigates the knowledge bias often observed in LLM-based judges that over-prioritize specific modalities. Evaluations across five diverse sensing benchmarks confirm that ConSensus outperforms the single-agent baseline by an average of 7.1% in accuracy. Furthermore, by employing a single-round hybrid fusion protocol, ConSensus achieves a 12.7× reduction in fusion token costs while matching the performance of state-of-the-art multi-agent debate methods.

Limitations

The scale of our experiments was constrained by the computational costs associated with multimodal sensing and multi-agent collaboration. In particular, processing heterogeneous sensor features through role-specified multi-agent collaboration incurs non-trivial inference overhead. To prioritize breadth of evaluation across diverse tasks, modalities, and baselines, we therefore conducted experiments on feasible subsets of each dataset rather than the full dataset. Future work should investigate the scalability of this framework to larger datasets and long-term sensing populations, as well as strategies for reducing inference overhead without sacrificing robustness.

This work is currently limited to classification tasks, as there is no established benchmark for evaluating LLM-based multimodal sensing across broader task types. We thus curated an evaluation suite by selecting datasets with objective ground-truth labels enable rigorous and reproducible accuracy-based evaluation. As a result, our evaluation does not cover human-centric reasoning or subjective judgment tasks. Extending the framework to such domains would require specialized data collection protocols and human annotations to assess the quality, coherence, and usefulness of generated reasoning.

This work focuses primarily on establishing modality-specific agents and the hybrid fusion protocol. As a result, we did not incorporate advanced prompting strategies, such as Self-Consistency (SC) or iterative Self-Refinement (SR), applied on top of ConSensus. Similarly, while multi-agent collaboration protocols such as ReConcile could potentially improve performance by integrating confidence signals into the semantic fusion process, we prioritized isolating and validating the effectiveness of the core protocol without introducing additional components. These advanced combinations represent promising future directions for enhancing the reasoning capabilities of both modality agents and fusion agents.

We utilized non-fine-tuned LLMs to demonstrate the generalizability of our framework. We expect that further specialization, such as fine-tuning agents on modality-specific data or integrating Retrieval-Augmented Generation (RAG) would further improve performance. By demonstrating that role specialization and structured data instructions alone yield significant gains, we establish a foundational step toward more complex multimodal systems. As sensor-specialized LLMs continue to emerge, ConSensus can serve as a guideline protocol for designing effective collaboration among heterogeneous, modality-aware agents.

Ethical Considerations

Potential Risks. ConSensus is a general multimodal sensing framework that can be applied to domains including health-related tasks. In high-stakes applications such as clinical decision support or mental health assessment, incorrect predictions or misleading model-generated reasoning may lead to inappropriate user actions. As LLMs may produce hallucinated or poorly calibrated interpretations, deploying such systems without adequate safeguards and human oversight may introduce safety risks. We emphasize that ConSensus is intended for research and exploratory use only. Any real-world deployment in safety-critical contexts should incorporate rigorous validation, regulatory compliance, and human-in-the-loop supervision. Further research is required to establish reliability guarantees, calibration mechanisms, and domain-specific safeguards before applying such systems to precision-sensitive applications.

Use of LLMs. We used ChatGPT and Gemini solely for language polishing of the manuscript. After the core implementation of our software, the Cursor AI assistant was used only for code formatting and cleanup.

References

Nimeesha Chan, Felix Parker, William C Bennett, Tianyi Wu, Mung Yao Jia, James Fackler MD, and Kimia Ghobadi. 2024. Leveraging LLMs for multimodal medical time series analysis. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Marquis de Condorcet. 1785. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, page 1785.

Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. 2022. Actionsense: A multimodal dataset and recording framework for human activities using wearable sensors in a kitchen environment. In *Advances in Neural Information Processing Systems*, volume 35, pages 13800–13813. Curran Associates, Inc.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Jinghan Du, Minghua Hu, and Weining Zhang. 2020. Missing data problem in the monitoring system: A review. *IEEE Sensors Journal*, 20(23):13984–13998.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

A. Ali Heydari, Ken Gu, Vidya Srinivas, Hong Yu, Zhihan Zhang, Yuwei Zhang, Akshay Paruchuri, Qian He, Hamid Palangi, Nova Hammerquist, Ahmed A. Metwally, Brent Winslow, Yubin Kim, Kumar Ayush, Yuzhe Yang, Girish Narayanswamy, Maxwell A. Xu, Jake Garrison, Amy Armento Lee, and 19 others. 2025. The anatomy of a personal health agent. *Preprint*, arXiv:2508.20148.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, zili wang, Steven Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Representation Learning*, volume 2024, pages 23247–23275.

Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2025. Llasa: A sensor-aware llm for natural language reasoning of human activity from imu data. *Preprint*, arXiv:2406.14498.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. 2000. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. In *Proceedings of the fifth Conference on Health, Inference, and Learning*, volume 248 of *Proceedings of Machine Learning Research*, pages 522–539. PMLR.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Tarlochan Singh Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Aalami, Filipe Barata, and Paul Schmiedmayer. 2025. Opentslm: Time-series language models for reasoning over multivariate medical text- and time-series data. *Preprint*, arXiv:2510.02410.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pages 51991–52008. Curran Associates, Inc.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large language models are few-shot health learners. *Preprint*, arXiv:2305.15525.

F Lotte, L Bougrain, A Cichocki, M Clerc, M Congedo, A Rakotomamonjy, and F Yger. 2018. A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob N. Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *CoRR*, abs/2408.06292.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Technical report, Meta Platforms, Inc.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Xiaomin Ouyang and Mani Srivastava. 2024. Llmsense: Harnessing llms for high-level reasoning over spatiotemporal sensor traces. In *2024 IEEE 3rd Workshop on Machine Learning on Edge in Sensor Systems (SenSys-ML)*, pages 9–14.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal deep learning for activity and context recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4).

Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109.

Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 400–408, New York, NY, USA. Association for Computing Machinery.

Leming Shen, Qiang Yang, Yuanqing Zheng, and Mo Li. 2025. *AutoIOT: LLM-Driven Automated Natural Language Programming for AIoT Applications*, page 468–482. Association for Computing Machinery, New York, NY, USA.

David Strömbäck, Sangxia Huang, and Valentin Radu. 2020. Mm-fit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 4(4).

Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*, 119:3–11. Deep Learning for Pattern Recognition.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024. Penetrative ai: Making llms comprehend the physical world. In *ACL (Findings)*, pages 7324–7341.

Xin Xu, Bei Zhang, Tingting Xu, and Junyi Tang. 2025. An effective and interpretable sleep stage classification approach using multi-domain electroencephalogram and electrooculogram features. *Bioengineering*, 12(3).

Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2024. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2219–2241, Miami, Florida, USA. Association for Computational Linguistics.

Xiaofan Yu, Lanxiang Hu, Benjamin Reichman, Dylan Chu, Rushil Chandrupatla, Xiyuan Zhang, Larry Heck, and Tajana S. Rosing. 2025. Sensorchat: Answering qualitative and quantitative questions during long-term multimodal sensor interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(3).

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. 2025. SensorLM: Learning the language of wearable sensors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Dataset	Task	#Classes	Sensor modalities
WESAD	Affective state recognition	3	ACC ECG EMG EDA TEMP RESP (Chest) ACC PPG EDA TEMP (Wrist)
SleepEDF	Sleep stage classification	5	EEG (Fpz-Cz) EEG (Pz-Oz) EOG EMG RESP
ActionSense	Kitchen activity recognition	5	ACC ANG EMG (Left arm), ACC ANG EMG (Right arm)
MMFit	Gym exercise recognition	10	ACC GYR × (Ear Left wrist Right wrist Waist) MAG (Waist) HR × (Left wrist Right wrist)
PAMAP2	Daily activity recognition	12	ACC GYR × (Hand Chest Ankle)

Table 3: Summary of multimodal sensing datasets and their respective sensor modalities.

Sensor	Extracted features
ACC GYR MAG ANG	Mean, std, and absolute integral for each axis (x, y, z) and magnitude. Peak frequency per axis.
ECG PPG	Heart rate (HR): Mean and std of HR derived from inter-beat intervals. HR variability (HRV): RMSSD, pNN50, TINN, and std. Frequency domain: Power in ULF (0.01–0.04 Hz), LF (0.04–0.15 Hz), HF (0.15–0.4 Hz), and UHF (0.4–1.0 Hz). Total power, LF/HF ratio, relative powers, and normalized LF/HF components.
EDA	Mean, std, min, max, slope, and dynamic range (5 Hz low-pass filtered). Tonic (SCL): Mean, std, and correlation with time. Phasic (SCR): Mean, std, event count, sum of magnitudes, total duration, and area under curve (AUC).
EMG	Chain 1 (high-pass): Mean, std, dynamic range, absolute integral, median, 10th/90th percentiles. Mean/median/peak frequency and spectral energy across seven bands (0–350 Hz). Chain 2 (50 Hz low-pass): Peak count, mean/std/sum of peak amplitudes, and normalized sum of amplitudes.
RESP	Inhalation/exhalation durations (mean, std, ratio), stretch, inspiration volume, respiration rate, and average cycle duration (0.1–0.35 Hz bandpass).
TEMP	Mean, std, min, max, slope, and dynamic range.
EEG	Frequency bands: Delta, theta, alpha, beta, spindle, K-complex, and sawtooth. Band features: Mean, std, variance, dynamic range, peak count, zero-crossing rate, variance of first-order difference, and absolute power (Welch's method). Ratios: Delta/theta, theta/alpha, alpha/beta, and (delta+theta)/(alpha+beta).
EOG	Time Domain: Mean, std, variance, dynamic range, zero-crossings, and first-order difference variance. Eye movements: Large movement count ($> 120\mu V$ within 1.5s) and difference variance (clean signal). Spectral: Slow (0.5–2 Hz) and rapid (2–5 Hz) power ratios relative to total power (0.5–30 Hz).

Table 4: Summary of features extracted for different sensor types.

A Dataset Details

Overview. Table 3 summarizes the tasks and sensor modalities for each dataset. The datasets include five to eleven modalities per task, covering 12 distinct sensor types collected from diverse devices and body locations, demonstrating broad diversity in sensing configurations.

Data Preprocessing. Following the protocols established in the original publications (Schmidt et al., 2018; Kemp et al., 2000; DelPreto et al., 2022; Strömbäck et al., 2020; Reiss and Stricker, 2012), raw sensor data were segmented using sliding windows. To prevent data leakage between few-shot examples and target samples, we employed non-overlapping windows (where step size equals window size). For WESAD, we focused on three affective states: baseline, stress, and amusement. For SleepEDF, we utilized the standard sleep stages (W, N1, N2, N3, and REM), excluding infrequent classes. In ActionSense, we grouped the labels into five high-level categories (spreading, peeling/slicing, jar operations, wiping, and tableware tasks) following the categorization in DelPreto et al. (2022); "pouring" was excluded due to insufficient samples. We utilized data from both the left and right armbands. For MMFit, we excluded data from the left smartphone due to sample sparsity. For PAMAP2, we selected 12 of the 18 original activity classes (e.g., walking, cycling, ironing) that provided sufficient sample density.

Feature Extraction. Features were extracted according to dataset-specific guidelines or established

Method	WESAD	SleepEDF	ActionSense	MMFit	PAMAP2	Avg.
Modality agents (single-round)						
+ Debate	0.907 ± 0.024	0.588 ± 0.031	0.607 ± 0.031	0.953 ± 0.012	0.551 ± 0.026	0.721 ± 0.025
+ MAD	0.847 ± 0.030	0.551 ± 0.030	0.591 ± 0.030	0.960 ± 0.011	**0.581** ± 0.026	0.706 ± 0.025
+ CMD	0.899 ± 0.025	0.575 ± 0.032	0.608 ± 0.031	0.964 ± 0.011	0.553 ± 0.026	0.720 ± 0.025
+ ReConcile	0.886 ± 0.026	0.557 ± 0.030	0.596 ± 0.031	0.944 ± 0.013	0.566 ± 0.026	0.710 ± 0.025
+ Semantic Fusion	0.825 ± 0.031	0.580 ± 0.031	0.605 ± 0.033	0.964 ± 0.011	0.559 ± 0.026	0.707 ± 0.026
+ Statistical Fusion	**0.927** ± 0.021	0.592 ± 0.032	0.597 ± 0.033	0.960 ± 0.011	0.534 ± 0.026	0.722 ± 0.026
ConSensus	0.880 ± 0.029	**0.600** ± 0.031	**0.611** ± 0.031	**0.967** ± 0.010	0.558 ± 0.026	**0.723** ± 0.026

Table 5: Comparison of ConSensus with multi-agent debate baselines under equal token cost (without iterative debate rounds). Results for the individual semantic and statistical fusion agents are also reported for reference. The best and second-best accuracies are highlighted in bold and underline, respectively.

signal processing methods (Schmidt et al., 2018; Xu et al., 2025), as summarized in Table 4. To evaluate system robustness against sensor failure, we simulated 10%, 30%, and 50% modality dropout rates by randomly masking sensor streams with zeros prior to feature extraction. Detailed task descriptions, class definitions, and feature extraction procedures were stored as metadata and dynamically parsed into the LLM prompts to provide environmental context.

Data Split. We split each dataset into an example set (for 1-shot in-context learning) and a test set to prevent data leakage. We adopt a within-subject split strategy, assuming each user provides a single data instance per class with minimal effort. Sensor data such as EEG exhibit substantial inter-subject domain shift (Lotte et al., 2018; Wang et al., 2019), which even remains an open challenge in the sensing community. Since LLMs are highly sensitive to in-context examples, cross-subject examples can incur significant performance variance. As our primary goal is to analyze *multimodal fusion behavior* under controlled conditions, we adopt within-subject calibration to isolate fusion effects from confounding domain-shift factors.

Data Subsampling. Due to the high token cost of using heavy sensor feature sequences with multiple LLMs (`gpt-oss-20B`, `gpt-oss-120B`, `Llama-3.1-8B-Instruct`, and `Llama-4-Scout-17B-16E`) and multi-agent baselines that require iterative debate rounds, we evaluate ConSensus on balanced dataset subsets. We sample 50 instances per class for WESAD, SleepEDF, and ActionSense (datasets with < 10 classes), and 30 instances per class for MMFit and PAMAP2 (datasets with ≥ 10 classes), resulting in test sets of 150–360 samples per task. This subsampling strategy is adopted solely to ensure computationally feasible comparison across methods under identical budgets, rather than as a requirement of the proposed framework.

We construct a single maximally large and diverse test split for each dataset to maximize sample coverage under limited computational budgets. Since ConSensus is training-free and deterministic (temperature$=0$ for all methods except Self-Consistency), the main source of randomness arises from dataset subsampling. We therefore report standard deviations using 1,000-iteration bootstrap resampling to estimate variance over samples. This evaluation protocol is adopted to prioritize broad benchmark coverage with multiple tasks and sensors, rather than to restrict statistical rigor.

B Multi-Agent Debate Performance without Iterative Rounds

In our main experiments, multi-agent debate baselines (Debate, MAD, CMD, ReConcile) employ two additional iterative rounds to refine initial predictions. These additional rounds incur substantially higher token cost than ConSensus. Without these iterative rounds, Debate and CMD reduce to majority voting, equivalent to our statistical fusion agent, while MAD reduces to an LLM judge, equivalent to our semantic fusion agent. ReConcile differs by applying confidence-weighted voting.

To compare performance under comparable token budgets with ConSensus, we evaluate all baselines without iterative debate rounds. Table 5 reports the results. Debate and CMD exhibit identical performance to statistical fusion, and MAD mirrors semantic fusion, as expected. ReConcile shows reduced accuracy (71.0%) when iterative refinement is removed, performing 1.3% lower than ConSensus. Overall, ConSensus consistently outperforms debate-based baselines under equal token cost, demonstrating that its hybrid fusion achieves superior accuracy without relying on expensive iterative deliberation.

C Prompt Templates

We provide the detailed prompts used in the single-agent baseline and ConSensus for the SleepEDF (Kemp et al., 2000) dataset. We present the `SYSTEM` and `USER` prompts for each specialized agent. To maintain brevity, recurring content or dataset-specific parameters are represented as placeholders (e.g., `<classes>`).

Single-Agent Baseline. The system prompt assigns the agent's role and details the task context. The user prompt provides a one-shot example and the multimodal features. It mandates a structured JSON output to facilitate automated parsing of the agent's reasoning.

> `System prompt`
>
> You are multimodal sensing agent that solves a sensing task. You have the following information about the task:
> **Task**: Classify the user's sleep stage: `<classes>`, based on physiological signals collected from wearable sensors.
> **Classes**: `<description of the classes>`
> You will receive sensor features from multiple modalities, and you have the following information about the modality: {`<modality 1>`: {"Data collection": `<data collection protocol>`, "Feature extraction": `<feature extraction methods>`}, `<modality 2>`: ..., *(repeated for all modalities)*}
>
> Your goal is to analyze the features and provide a reasoned answer using your knowledge.

> `User prompt`
>
> You have received sensor features from multiple modalities:
> **Examples:**
> Sensor values might not always align with your inherent knowledge due to differences in data collection or processing. So, we included a few labeled examples to help your interpretation:
> *Example of* `<class 1>`:
> - `<feature name>`: `<value>`
> - *(repeated for all modality features)*
> *Example of* `<class 2>`: ...
> *(repeated for all examples)*
> **Current sample features**: ...
>
> Please provide your answer for the task among `<classes>` and the reasoning for your answer. Note that the sensor features might be wrong due to the data collection or processing. You can evaluate the quality of the features by checking the examples you have.
>
> Respond in the following strict JSON format: {"REASON": "<Reasoning for the answer>", "ANSWER": "<Answer among `<classes>`>"}
> Do not include any additional text outside of the JSON.

Modality Agents. We provide a modality agent prompt for `EEG-Pz-Oz` in the SleepEDF dataset. This agent has a modality-specific role while using the same task template as the single-agent baseline. It follows the same user prompt structure as the single-agent baseline; the only difference is that it includes only the `EEG-Pz-Oz` features instead of the full multimodal feature set.

> `System prompt`
>
> You are EEG-Pz-Oz agent that solves a sensing task.
> `<task description prompt>`

> `User prompt`
>
> You have received sensor features from EEG-Pz-Oz modality:
> `<modality-specific example prompt>`
> `<modality-specific sample prompt>`
>
> `<instruction prompt>`
> `<formatting prompt>`

Semantic Fusion Agent. We assign the role of a fusion agent and instruct it to aggregate multiple outputs from the modality agents to solve the task. The outputs from all modality agents are merged and passed to the prompt. It is then instructed to generate a final output based on its own knowledge and expertise.

`System prompt`

You are a fusion agent that solves a multimodal sensing task based on interpretations from multiple sensing agents.
`<task description prompt>`

You will receive reasonings and answers from multiple agents based on their interpretations of different modalities. Your goal is to provide a final reasoned answer for the task.

`User prompt`

You have received responses from multiple sensing agents:
{"EEG-Fpz-Cz": `<modality agent output>`, "EMG-submental": ..., *(repeated for all modality agents)*}

Using your own knowledge and expertise, analyze the reasonings and answers and provide a final reasoned answer.
`<formatting prompt>`

Statistical Fusion Agent. The statistical fusion agent shares the same system prompt as the semantic fusion agent but utilizes a distinct user prompt. It is provided with the `<majority-voted answer>` and tasked with justifying this consensus by identifying potential failure modes in the dissenting modality agents.

`User prompt`

You have received responses from multiple sensing agents:
`<responses from modality agents>`

You are on the side that the correct answer is `<majority-voted answer>` which is the majority answer. Based on the given reasonings and answers, provide a final reasoned answer for the task.

For agents who provided different answers from `<majority-voted answer>`, explain why their reasoning is likely affected by noise or unreliable signal interpretation. Stay consistent with the position that the correct answer is likely `<majority-voted answer>`.
`<formatting prompt>`

Hybrid Fusion Agent. The hybrid fusion agent is guided by a system prompt that assigns it the role of a coordinator for reconciling the outputs of the fusion agents. It is provided with (i) the modality agents' outputs and (ii) the responses from both semantic and statistical fusion agents. Its task is to evaluate these responses and provide the final consolidated output.

`System prompt`

You are a coordinator agent that solves a multimodal sensing task based on interpretations from multiple sensing agents.
`<task description prompt>`

You will receive reasonings and answers from multiple agents, each interpreting different modalities, as well as two fusion agents that have already aggregated these responses in different ways. Your goal is to coordinate the responses from the fusion agents and provide a final reasoned answer for the task.

`User prompt`

You have received the following information.
1. Original responses from individual modality agents:
`<responses from modality agents>`

2. Responses from fusion agents:
`<response from the semantic fusion agent>`
`<response from the statistical fusion agent>`

Your task is to determine the responses from the fusion agents and provide a final reasoned answer for the task.
`<formatting prompt>`

D Qualitative Analysis

We present outputs from the single-agent baseline and ConSensus to illustrate the collaborative decision-making process, using examples from the SleepEDF dataset.

Single-Agent Baseline. In this example, the ground-truth label is REM, but the single-agent baseline incorrectly predicts N2. While the agent performs plausible reasoning on the spectral features, it fails to integrate the subtle respiratory cues that output REM from light sleep.

`LLM response` (Sinlge-agent)

{"REASON": "The ratio of (delta+theta) to (alpha+beta) for both Fpz-Cz (9.34) and Pz-Oz (3.75) falls between the typical values for wakefulness ($\approx$ 6.1 and 1.5) and stage N2 ($\approx$ 10.6 and 4.5). The delta power (86.75) is higher than in N1 but much lower than in deep sleep (N3). Alpha and beta powers are lower than in N1/N2 but higher than in wakefulness, indicating a transition toward sleep. The k-complex power and EMG activity are low, suggesting a lighter sleep stage. EOG shows many large eye movements and a high slow-movement ratio, which is more consistent with wake or N1 rather than REM. Overall, the feature pattern best matches stage N2, the intermediate sleep stage.", "ANSWER": "N2"}

However, when isolating the respiration data via a specialized modality agent, the correct stage is identified:

`LLM response` (Respiration modality agent)

{"REASON": "The current sample shows an inhale duration longer than exhale (ratio 1.39) and an inhale mean of 2.49 s versus exhale mean of 1.79 s, which aligns with the REM example where the inhale is longer than the exhale (ratio 1.33). Other REM characteristics—such as a relatively high respiration rate ($\sim$13 breaths/min), moderate duration ($\sim$4.6 s), and moderate stretch—are also reflected in the current data. In contrast, the other stages (W, N1, N2, N3) all have inhale/exhale ratios below 1 and different mean durations. Therefore, the most consistent match is REM.", "ANSWER": "REM"}

Semantic Fusion Agent. In the previous example, while several modality agents provides incorrect classifications, the EEG-Pz-Oz and respiration agents correctly identify the REM stage. The semantic fusion agent successfully resolves this cross-modal conflict by prioritizing the high-fidelity evidence correct modality agents.

`LLM response` (Semantic fusion agent)

{"REASON": "The most reliable indicator of sleep stage is the EEG. While the frontal channel (Fpz-Cz) shows a high (delta+theta)/(alpha+beta) ratio typical of N2, the posterior channel (Pz-Oz) displays a ratio (3.75) that lies between the typical REM value ($\approx$2.6) and the N2 value ($\approx$4.5), and its other features (lower alpha/beta, lower variance) are more consistent with REM than with deeper stages. The respiratory pattern shows an inhale/exhale ratio $>$1 and a respiration rate ($\sim$13 breaths/min) characteristic of REM. EMG amplitude is low compared with wakefulness, which aligns with REM muscle atonia, although it is not as low as would be seen in deep sleep. EOG indicates a high slow-movement ratio, but this can be influenced by sensor noise and is not a definitive marker for N3. Considering the stronger evidence from EEG (posterior channel) and respiration, the overall pattern points to REM sleep. Therefore the most plausible classification is REM.", "ANSWER": "REM"}

However, the semantic fusion agent is susceptible to prior-knowledge bias, where it may over-rely on a single modality output even when it is incorrect. For example, in an N1 stage sample, we observed a case where the respiration agent incorrectly suggested N2; despite the majority of other agents correctly identifying N1, the semantic agent erroneously followed the respiration agent.

`LLM response` (Semantic fusion agent)

{"REASON": "The most reliable signals for sleep staging are the EEG and respiration measures. Both EEG channels report low-amplitude, low-power activity, which is inconsistent with deep sleep (N3) that would show high-amplitude slow waves, and also inconsistent with REM, which would typically show characteristic sawtooth activity and higher dynamic range. The EOG channel's lack of large eye movements rules out REM and is compatible with both light and intermediate stages, but does not distinguish N1 from N2. The respiration data shows a moderate stretch ($\approx$ 2.2×10^9L), a respiration rate of 13.4

bpm, and exhale/inhale durations around 2.2 s—values that fall squarely within the typical range for stage N2 and are less compatible with the lower volume, longer duration pattern of N3 or the shorter, higher-rate pattern of REM. The EMG signal is intermediate between wakefulness and deeper stages, again pointing to a transitional stage rather than deep sleep. Given that the majority of modalities (EEG, respiration, EMG) lean toward an intermediate sleep stage and the EOG does not provide evidence for REM, the most consistent classification is stage N2.", "ANSWER": "N2"}

Statistical Fusion Agent. To mitigate semantic errors, the statistical fusion agent provides a complementary perspective by anchoring its reasoning to the majority consensus. In the aforementioned failure case, it identifies that the majority of modality agents point to N1 and provides a justification for why dissenting agents are likely erroneous.

`LLM response` (Statistical fusion agent)

{"REASON": "The two modalities that are most directly linked to sleep stage classification—central EEG (Fpz-Cz) and submental EMG—both point to light sleep (N1). The Fpz-Cz feature vector shows low-amplitude, low-power characteristics (high variance relative to N1, low delta power, low delta/theta ratio, and an alpha/beta ratio close to the N1 mean) that match the prototypical N1 profile far better than any other stage. The submental EMG is in the intermediate activity range between wakefulness and deeper stages, with mean and standard deviation values, percentiles, and peak statistics that align with published N1 values. These two signals are the gold standards for distinguishing wake, N1, N2, N3, and REM, so their agreement strongly supports an N1 classification. The other modalities provide weaker or conflicting evidence. The EOG horizontal channel, while showing low rapid-movement power and high slow-movement power, also exhibits an unusually high dynamic range and a very low difference variance, which are typical of measurement artifacts or electrode misplacement. Because eye-movement activity is only one component of REM and N3, relying on it alone can misclassify light sleep as deep sleep. The EEG-Pz-Oz channel indicates REM based on low dynamic range and sawtooth power, yet its (delta+theta)/(alpha+beta) ratio is intermediate and the alpha/beta ratio is higher than the REM prototype, suggesting noise or cross-talk from the adjacent frontal channel. Finally, the respiratory signal suggests N2 due to moderate breath duration and volume, but respiration alone is a poor discriminator between N2 and N3 and is highly variable across individuals. Given the central EEG and EMG consensus and the known unreliability of single-modality cues in isolation, the most reliable conclusion is that the subject is in the N1 stage.", "ANSWER": "N1"}

Hybrid Fusion Agent. The hybrid fusion agent then evaluates these competing reasoning paths. In the previous example, it follows the statistical consensus.

`LLM response` (Hybrid fusion agent)

{"REASON": "The most reliable evidence comes from the central EEG (Fpz-Cz) and submental EMG, both of which are the gold standards for distinguishing sleep stages. The Fpz-Cz channel shows low-amplitude, low-power activity with variance and delta power values that match the prototypical N1 profile far better than any other stage. The submental EMG is in the intermediate activity range between wakefulness and deeper stages, with mean, standard deviation, and peak statistics that align closely with published N1 values. These two signals agree on light sleep (N1). Other modalities provide conflicting or weaker cues: the Pz-Oz channel suggests REM but its intermediate (delta+theta)/(alpha+beta) ratio and higher alpha/beta ratio point to noise; the EOG horizontal channel indicates minimal rapid eye movements, which is typical of both N1 and N2 and not decisive; the respiration signal is consistent with N2 but respiration alone is a poor discriminator. Given the consensus of the most reliable modalities and the known unreliability of single-modality cues in isolation, the most robust conclusion is that the subject is in the N1 stage.", "ANSWER":"N1"}